09/462845
GC382-PCT
PCT/US 98/14647

Europäisches Patentamt | European Patent Office | Office européen des brevets

31 AUG 1998

17 SEP 1998

Bescheinigung | Certificate | Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

97305232.7

PRIORITY DOCUMENT

Der Präsident des Europäischen Patentamts:
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

M.B. RIJLING

DEN HAAG, DEN
THE HAGUE, 17/08/98
LA HAYE, LE

Europäisches Patentamt | European Patent Office | Office européen des brevets

Blatt 2 der Bescheinigung
Sheet 2 of the certificate
Page 2 de l'attestation

Anmeldung Nr.: Application no.: Demande n°: 97305232.7

Anmeldetag: Date of filing: Date de dépôt: 15/07/97

Anmelder: Applicant(s): Demandeur(s):
Genencor International B.V.
2600 AP Delft
NETHERLANDS

Bezeichnung der Erfindung: Title of the invention: Titre de l'invention:
Proteases from gram-positive organisms

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

Staat State Pays	Tag Date Date	Aktenzeichen: File no. Numéro de dépôt:

Internationale Patentklassifikation
International Patent classification
Classification internationale des brevets
/

Am Anmeldetag benannte Vertragstaaten
Contracting states designated at date of filing: AT/BE/CH/DE/DK/ES/FI/FR/GB/GR/IE/IT/LI/LU/MC/NL/PT/SE
Etats contractants désignés lors du dépôt

Bemerkungen
Remarks
Remarques

PATENT
Docket No. GC 382

PROTEASES FROM GRAM-POSITIVE ORGANISMS

FIELD OF THE INVENTION

The present invention relates to serine proteases derived from gram-positive microorganisms. The present invention provides nucleic acid and amino acid sequences of serine protease 1, 2, 3, 4 and 5 identified in *Bacillus*. The present invention also provides methods for the production of serine protease 1, 2, 3, 4 and 5 in host cells as well as the production of heterologous proteins in a host cell having a mutation or deletion of part or all of at least one of the serine proteases of the present invention.

BACKGROUND OF THE INVENTION

Gram-positive microorganisms, such as members of the group *Bacillus*, have been used for large-scale industrial fermentation due, in part, to their ability to secrete their fermentation products into the culture media. In gram-positive bacteria, secreted proteins are exported across a cell membrane and a cell wall, and then are subsequently released into the external media usually maintaining their native conformation.

Various gram-positive microorganisms are known to secrete extracellular and/or intracellular protease at some stage in their life cycles. Many proteases are produced in large quantities for industrial purposes. A negative aspect of the presence of proteases in gram-positive organisms is their contribution to the overall degradation of secreted heterologous or foreign proteins.

The classification of proteases found in microorganisms is based on their catalytic mechanism which results in four groups: the serine proteases; metalloproteases; cysteine proteases; and aspartic proteases. These categories can be distinguished by their sensitivity to various inhibitors. For example, the serine proteases are inhibited by phenylmethylsulfonylfluoride (PMSF) and diisopropylfluorophosphate (DIFP); the metalloproteases by chelating agents; the cysteine enzymes by iodoacetamide and heavy metals and the aspartic proteases by pepstatin. The serine proteases have alkaline pH optima, the metalloproteases are optimally active around neutrality, and the cysteine and aspartic enzymes have

acidic pH optima (Biotechnology Handbooks, *Bacillus*. vol. 2, edited by Harwood, 1989 Plenum Press, New York).

Proteolytic enzymes that are dependent upon a serine residue for catalytic activity are called serine proteases. As described in Methods in Enzymology, vol. 244, Academic Press, Inc. 1994, page 21, serine proteases of the family S9 have the catalytic residue triad "Ser-Asp-His with conservation of amino acids around them.

SUMMARY OF THE INVENTION

The present invention relates to the unexpected discovery of five heretofore unknown or unrecognized S9 type serine proteases found in uncharacterized translated genomic nucleic acid sequences of *Bacillus subtilis*, designated herein as SP1, SP2, SP3, SP4 and SP5 having the nucleic acid and amino acid as shown in the Figures. The present invention is based, in part, upon the presence the amino acid triad S-D-H in the five serine proteases, as well as amino acid conservation around the triad. The present invention is also based in part upon the heretofore uncharacterized or unrecognized overall amino acid relatedness that SP1, SP2, SP3, SP4 and SP5 have with the serine protease dipeptidyl-amino peptidase B from yeast (DAP) and with each other.

The present invention provides isolated polynucleotide and amino acid sequences for SP1, SP2, SP3, SP4 and SP5. Due to the degeneracy of the genetic code, the present invention encompasses any nucleic acid sequence that encodes the SP1, SP2, SP3, SP4 and SP5 deduced amino acid sequences shown in Figures 2A-2B-Figure 6, respectively.

The present invention encompasses amino acid variations of *B.subtilis* SP1, SP2, SP3, SP4 and SP5 amino acids disclosed herein that have proteolytic activity. *B. subtilis* SP1, SP2, SP3, SP4 and SP5 as well as proteolytically active amino acid variations, thereof have application in cleaning compositions. The present invention also encompasses amino acid variations or derivatives of SP1, SP2, SP3, SP4 and SP5 that do not have the characteristic proteolytic activity as long as the nucleic acid sequences encoding such variations or derivatives would have sufficient 5' and 3' coding regions to be capable of integration into a gram-positive organism genome. Such variants would have applications in gram-positive expression systems where it is desirable to delete, mutate, alter or otherwise incapacitate the naturally occurring serine protease in order to diminish or delete its proteolytic activity. Such an

expression system would have the advantage of allowing for greater yields of recombinant heterologous proteins or polypeptides.

The present invention provides methods for detecting gram positive microorganism homologs of *B. subtilis* SP1, SP2, SP3, SP4 and SP5 that comprises hybridizing part or all of the nucleic acid encoding *B. subtilis* SP1, SP2, SP3, SP4 or SP45 with nucleic acid derived from gram-positive organisms, either of genomic or cDNA origin. In one embodiment, the gram-positive microorganism is selected from the group consisting of *B. licheniformis, B. lentus, B. brevis, B. stearothermophilus, B. alkalophilus, B. amyloliquefaciens, B. coagulans, B. circulans, B. lautus and Bacillus thuringiensis.*

The production of desired heterologous proteins or polypeptides in gram-positive microorganisms may be hindered by the presence of one or more proteases which degrade the produced heterologous protein or polypeptide. One advantage of the present invention is that it provides methods and expression systems which can be used to prevent that degradation, thereby enhancing yields of the desired heterologous protein or polypeptide.

Thus, in another aspect, the present invention provides a gram-positive microorganism having a mutation or deletion of part or all of the gene encoding SP1 and/or SP2 and/or SP3 and/or SP4 and/or SP5 which results in inactivation of their proteolytic activity, either alone or in combination with mutations in other proteases, such as apr, npr, epr, mpr for example, or other proteases known to those of skill in the art. In one embodiment of the present invention, the gram-positive organism is a member of the genus *Bacillus*. In another embodiment, the *Bacillus* is *Bacillus subtilis*.

In yet another aspect, the gram-positive host is genetically engineered to produce a desired protein. In one embodiment of the present invention, the desired protein is heterologous to the gram-positive host cell. In another embodiment, the desired protein is homologous to the host cell. The present invention encompasses a gram-positive host cell having a deletion or interruption of the nucleic acid encoding the naturally occurring homologous protein, such as a protease, and having nucleic acid encoding the homologous protein re-introduced in a recombinant form. In another embodiment, the host cell produces the homologous protein. Accordingly, the present invention also provides methods and expression systems for reducing degradation of heterologous proteins produced in gram-positive microorganisms. The gram-positive microorganism may be normally sporulating or non-sporulating.

In a further aspect of the present invention, gram-positive SP1, SP2, SP3, SP4 or SP5 is produced on an industrial fermentation scale in a microbial host expression system. In another aspect, isolated and purified recombinant SP1, SP2, SP3, SP4 or SP5 is used in compositions of matter intended for cleaning purposes, such as detergents. Accordingly, the present invention provides a cleaning composition comprising one or more of a gram-positive serine protease selected from the group consisting of SP1, SP2, SP3, SP4 and SP45. The serine protease may be used alone or in combination with other enzymes and/or mediators or enhancers.

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1A-1C shows the DNA and deduced amino acid sequence for SP1 (YUXL).

Figure 2A-2B show an amino acid alignment between DAP (dap2_yeast) and SP1 (YUXL). For Figures 2A-2B, 3 and 4, the amino acid triad S-D-H is indicated.

Figure 3 shows an amino acid alignment between SP1 (YUXL) and SP2 (YTMA).

Figure 4 shows and amino acid alignment between SP1 (YUXL) and SP3 (YITV).

Figure 5 shows and amino acid alignment between SP1 (YUXL) and SP4 (YQKD).

Figure 6 shows and amino acid alignment between SP1 (YUXL) and SP5 (CAH).

Figures 7A-7B shows the DNA and deduced amino acid sequence for SP2 (YTMA).

Figures 8A-8B shows the DNA and deduced amino acid sequence for SP3 (YITV).

Figures 9A-9B shows the DNA and deduced amino acid sequence for SP4 (YQKD).

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Definitions

As used herein, the genus *Bacillus* includes all members known to those of skill in the art, including but not limited to *B. subtilis, B. licheniformis, B. lentus, B.*

brevis, B. stearothermophilus, B. alkalophilus, B. amyloliquefaciens, B. coagulans, B. ciculans, B. lautus and B. thuringiensis.

The present invention encompasses novel SP1, SP2, SP3, SP4 and SP5 from gram positive organisms. In a preferred embodiment, the gram-positive organisms is a *Bacillus*. In another preferred embodiment, the gram-positive organism is *Bacillus subtilis*. As used herein, "*B.subtilis* SP1 (YuxL) refers to the DNA and deduced amino acid sequence shown in Figures 1A-1C and Figures 2A-2B; SP2 (YtmA) refers to the DNA and deduced amino acid sequence shown in Figures 7A-7B and Figure 3; SP3 (YitV) refers to the DNA and deduced amino acid sequence shown in Figures 8A-8B and Figure 4; SP4 (YqkD) refers to the DNA and deduced amino acid sequence shown in Figures 9A-9B and Figure 5; and SP5 (CAH) refers to the deduced amino acid sequence shown in Figure 6. The present invention encompasses amino acid variations of the *B.subtilis* amino acid sequences of SP1, SP2, SP3, SP4 and SP5 that have proteolytic activity. Such proteolytic amino acid variants can be used in cleaning compositions. The present invention also encompasses *B. subtilis* amino acid variations or derivatives that are not proteolytically active. DNA encoding such variants can be used in methods designed to delete or mutate the naturally occurring host cell SP1, SP2, SP3, SP4 and SP5.

As used herein, "nucleic acid" refers to a nucleotide or polynucleotide sequence, and fragments or portions thereof, and to DNA or RNA of genomic or synthetic origin which may be double-stranded or single-stranded, whether representing the sense or antisense strand. As used herein "amino acid" refers to peptide or protein sequences or portions thereof. A "polynucleotide homolog" as used herein refers to a novel gram-positive microorganism polynucleotide that has at least 80%, at least 90% and at least 95% identity to *B.subtilis* SP1, SP2, SP3, SP4 or SP5, or which is capable of hybridizing to *B.subtilis* SP1, SP2, SP3, SP4 or SP5 under conditions of high stringency and which encodes an amino acid sequence having serine protease activity.

The terms "isolated" or "purified" as used herein refer to a nucleic acid or amino acid that is removed from at least one component with which it is naturally associated.

As used herein, the term "heterologous protein" refers to a protein or polypeptide that does not naturally occur in a gram-positive host cell. Examples of heterologous proteins include enzymes such as hydrolases including proteases, cellulases, amylases, carbohydrases, and lipases; isomerases such as racemases,

epimerases, tautomerases, or mutases; transferases, kinases and phophatases. The heterologous gene may encode therapeutically significant proteins or peptides, such as growth factors, cytokines, ligands, receptors and inhibitors, as well as vaccines and antibodies. The gene may encode commercially important industrial proteins or peptides, such as proteases, carbohydrases such as amylases and glucoamylases, cellulases, oxidases and lipases. The gene of interest may be a naturally occurring gene, a mutated gene or a synthetic gene.

The term "homologous protein" refers to a protein or polypeptide native or naturally occurring in a gram-positive host cell. The invention includes host cells producing the homologous protein via recombinant DNA technology. The present invention encompasses a gram-positive host cell having a deletion or interruption of the nucleic acid encoding the naturally occurring homologous protein, such as a protease, and having nucleic acid encoding the homologous protein re-introduced in a recombinant form. In another embodiment, the host cell produces the homologous protein.

As used herein, the term "overexpressing" when referring to the production of a protein in a host cell means that the protein is produced in greater amounts than its production in its naturally occurring environment.

As used herein, the phrase "proteolytic activity" refers to a protein that is able to hydrolyze a peptide bond. Enzymes having proteolytic activity are described in Enzyme Nomenclature. 1992, edited Webb Academic Press, Inc.

Detailed Description of the Preferred Embodiments

The unexpected discovery of the serine proteases SP1, SP2, SP3, SP4 and SP5 in *B.subtilis* provides a basis for producing host cells, expression methods and systems which can be used to prevent the degradation of recombinantly produced heterologous proteins. In a preferred embodiment, the host cell is a gram-positive host cell that has a reduction or mutation in the naturally occurring serine protease said mutation resulting in the complete deletion or inactivation of the production by the host cell of the proteolytic serine protease gene product. In another embodiment of the present invention, the host cell is additionally genetically engineered to produced a desired protein or polypeptide.

It may also be desired to genetically engineer host cells of any type to produce a gram-positive serine protease SP1, SP2, SP3, SP4 or SP5. Such host

cells are used in large scale fermentation to produce large quantities of the serine protease which may be isolated or purified and used in cleaning products, such as detergents.

I. Serine protease Sequences

The SP1, SP2, SP3 and SP4 polynucleotides having the sequences as shown in the Figures encode the *Bacillus subtilis* serine SP1, SP2, SP3, and SP4. As will be understood by the skilled artisan, due to the degeneracy of the genetic code, a variety of polynucleotides can encode the *Bacillus* SP1, SP2, SP3, SP4 and SP5. The present invention encompasses all such polynucleotides.

The present invention encompasses novel SP1, SP2, SP3, SP4 and SP5 polynucleotide homologs encoding gram-positive microorganism serine proteases SP1, SP2, SP3, SP4 and SP5, respectively, which have at least 80%, or at least 90% or at least 95% identity to *B.subtilis* as long as the homolog encodes a protein that has proteolytic activity.

Gram-positive polynucleotide homologs of *B.subtilis* SP1, SP2, SP3, SP4 or SP5 may be obtained by standard procedures known in the art from, for example, cloned DNA (*e.g.*, a DNA "library"), genomic DNA libraries, by chemical synthesis once identified, by cDNA cloning, or by the cloning of genomic DNA, or fragments thereof, purified from a desired cell. (*See*, for example, Sambrook *et al.*, 1989, Molecular Cloning, A Laboratory Manual, 2d Ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York; Glover, D.M. (ed.), 1985, DNA Cloning: A Practical Approach, MRL Press, Ltd., Oxford, U.K. Vol. I, II.) A preferred source is from genomic DNA. Nucleic acid sequences derived from genomic DNA may contain regulatory regions in addition to coding regions. Whatever the source, the isolated serine protease gene should be molecularly cloned into a suitable vector for propagation of the gene.

In the molecular cloning of the gene from genomic DNA, DNA fragments are generated, some of which will encode the desired gene. The DNA may be cleaved at specific sites using various restriction enzymes. Alternatively, one may use DNAse in the presence of manganese to fragment the DNA, or the DNA can be physically sheared, as for example, by sonication. The linear DNA fragments can then be separated according to size by standard techniques, including but not limited to, agarose and polyacrylamide gel electrophoresis and column chromatography.

Once the DNA fragments are generated, identification of the specific DNA fragment containing the SP1, SP2, SP3, SP4 or SP5 may be accomplished in a number of ways. For example, a *B.subtilis* SP1, SP2, SP3, SP4 or SP5 gene of the

present invention or its specific RNA, or a fragment thereof, such as a probe or primer, may be isolated and labeled and then used in hybridization assays to detect a gram-positive SP1, SP2, SP3, SP4 or SP5 gene. (Benton, W. and Davis, R., 1977, Science 196:180; Grunstein, M. And Hogness, D., 1975, Proc. Natl. Acad. Sci. USA 72:3961). Those DNA fragments sharing substantial sequence similarity to the probe will hybridize under stringent conditions.

Accordingly, the present invention provides a method for the detection of gram-positive SP1, SP2, SP3, SP4 or SP5 polynucleotide homologs which comprises hybridizing part or all of a nucleic acid sequence of *B. subtilis* SP1, SP2, SP3, SP4 or SP5 with gram-positive microorganism nucleic acid of either genomic or cDNA origin.

Also included within the scope of the present invention are gram-positive microorganism polynucleotide sequences that are capable of hybridizing to the nucleotide sequence of *B.subtilis* SP1, SP2, SP3, SP4 or SP5 under conditions of intermediate to maximal stringency. Hybridization conditions are based on the melting temperature (Tm) of the nucleic acid binding complex, as taught in Berger and Kimmel (1987, Guide to Molecular Cloning Techniques, Methods in Enzymology, Vol 152, Academic Press, San Diego CA) incorporated herein by reference, and confer a defined "stringency" as explained below.

"Maximum stringency" typically occurs at about Tm-5°C (5°C below the Tm of the probe); "high stringency" at about 5°C to 10°C below Tm; "intermediate stringency" at about 10°C to 20°C below Tm; and "low stringency" at about 20°C to 25°C below Tm. As will be understood by those of skill in the art, a maximum stringency hybridization can be used to identify or detect identical polynucleotide sequences while an intermediate or low stringency hybridization can be used to identify or detect polynucleotide sequence homologs.

The term "hybridization" as used herein shall include "the process by which a strand of nucleic acid joins with a complementary strand through base pairing" (Coombs J (1994) Dictionary of Biotechnology, Stockton Press, New York NY).

The process of amplification as carried out in polymerase chain reaction (PCR) technologies is described in Dieffenbach CW and GS Dveksler (1995, PCR Primer, a Laboratory Manual, Cold Spring Harbor Press, Plainview NY). A nucleic acid sequence of at least about 10 nucleotides and as many as about 60 nucleotides from *B. subtilis* SP1, SP2, SP3, SP4 or SP5 preferably about 12 to 30 nucleotides, and more preferably about 20-25 nucleotides can be used as a probe or PCR primer.

The *B.subtilis* amino acid sequences SP1, SP2, SP3, SP4 and SP5 (shown in Figures 2A-2B through Figure 6) were identified via a FASTA search of *Bacillus subtilis* genomic nucleic acid sequences. *B. subtilis* SP1 (YuxL) was identified by its structural homology to the serine protease DAP classified as an S9 type serine protease, designated in Figures 2A-2B as "dap2_yeast". As shown in Figures 2A-2B, SP1 has the amino acid dyad "S-D-H" indicated. Conservation of amino acids around each residue is noted in Figures 2A-2B through Figure 6. SP2 (YtmA); SP3 (YitV); SP4 (YqkD0 and SP5 (CAH) were identified upon by their structural and overall amino acid homology to SP1 (YuxL). SP1 and SP4 were described in Parsot and Kebayashi, respectively, but were not characterized as serine proteases or serine proteases of the S9 family.

II. Expression Systems

The present invention provides host cells, expression methods and systems for the enhanced production and secretion of desired heterologous or homologous proteins in gram-positive microorganisms. In one embodiment, a host cell is genetically engineered to have a deletion or mutation in the gene encoding a gram-positive SP1, SP2, SP3, SP4 or SP5 such that the respective activity is deleted. In an alternative embodiment of the present invention, a gram-positive microorganism is genetically engineered to produce a serine protease of the present invention.

Inactivation of a gram-positive serine protease in a host cell

Producing an expression host cell incapable of producing the naturally occurring serine protease necessitates the replacement and/or inactivation of the naturally occurring gene from the genome of the host cell. In a preferred embodiment, the mutation is a non-reverting mutation.

One method for mutating nucleic acid encoding a gram-positive serine protease is to clone the nucleic acid or part thereof, modify the nucleic acid by site directed mutagenesis and reintroduce the mutated nucleic acid into the cell on a plasmid. By homologous recombination, the mutated gene may be introduced into the chromosome. In the parent host cell, the result is that the naturally occurring nucleic acid and the mutated nucleic acid are located in tandem on the chromosome. After a second recombination, the modified sequence is left in the chromosome having thereby effectively introduced the mutation into the chromosomal gene for progeny of the parent host cell.

Another method for inactivating the serine protease proteolytic activity is through deleting the chromosomal gene copy. In a preferred embodiment, the entire gene is deleted, the deletion occurring in such as way as to make reversion impossible. In another preferred embodiment, a partial deletion is produced, provided that the nucleic acid sequence left in the chromosome is too short for homologous recombination with a plasmid encoded serine protease gene. In another preferred embodiment, nucleic acid encoding the catalytic amino acid residues are deleted.

Deletion of the naturally occurring gram-positive microorganism serine protease can be carried out as follows. A serine protease gene including its 5' and 3' regions is isolated and inserted into a cloning vector. The coding region of the serine protease gene is deleted form the vector *in vitro*, leaving behind a sufficient amount of the 5' and 3' flanking sequences to provide for homologous recombination with the naturally occurring gene in the parent host cell. The vector is then transformed into the gram-positive host cell. The vector integrates into the chromosome via homologous recombination in the flanking regions. This method leads to a gram-positive strain in which the protease gene has been deleted.

The vector used in an integration method is preferably a plasmid. A selectable marker may be included to allow for ease of identification of desired recombinant microorgansims. Additionally, as will be appreciated by one of skill in the art, the vector is preferably one which can be selectively integrated into the chromosome. This can be achieved by introducing an inducible origin of replication, for example, a temperature sensitive origin into the plasmid. By growing the transformants at a temperature to which the origin of replication is sensitive, the replication function of the plasmid is inactivated, thereby providing a means for selection of chromosomal integrants. Integrants may be selected for growth at high temperatures in the presence of the selectable marker, such as an antibiotic. Integration mechanisms are described in WO 88/06623.

Integration by the Campbell-type mechanism can take place in the 5' flanking region of the protease gene, resulting in a protease positive strain carrying the entire plasmid vector in the chromosome in the serine protease locus. Since illegitimate recombination will give different results it will be necessary to determine whether the complete gene has been deleted, such as through nucleic acid sequencing or restriction maps.

Another method of inactivating the naturally occurring serine protease gene is to mutagenize the chromosomal gene copy by transforming a gram-positive microorganism with oligonucleotides which are mutagenic. Alternatively, the chromosomal serine protease gene can be replaced with a mutant gene by homologous recombination.

The present invention encompasses host cells having additional protease deletions or mutations, such as deletions or mutations in apr, npr, epr, mpr and others known to those of skill in the art.

III. Production of Serine protease

For production of serine protease in a host cell, an expression vector comprising at least one copy of nucleic acid encoding a gram-positive microorganism SP1, SP2, SP3, SP4 or SP5, and preferably comprising multiple copies, is transformed into the host cell under conditions suitable for expression of the serine protease. In accordance with the present invention, polynucleotides which encode a gram-positive microorganism SP1, SP2, SP3, SP4 or SP5, or fragments thereof, or fusion proteins or polynucleotide homolog sequences that encode amino acid variants of B. SP1, SP2, SP3, SP4 or SP5, may be used to generate recombinant DNA molecules that direct their expression in host cells. In a preferred embodiment, the gram-positive host cell belongs to the genus *Bacillus*. In another preferred embodiment, the gram positive host cell is *B. subtilis*.

As will be understood by those of skill in the art, it may be advantageous to produce polynucleotide sequences possessing non-naturally occurring codons. Codons preferred by a particular gram-positive host cell (Murray E et al (1989) Nuc Acids Res 17:477-508) can be selected, for example, to increase the rate of expression or to produce recombinant RNA transcripts having desirable properties, such as a longer half-life, than transcripts produced from naturally occurring sequence.

Altered SP1, SP2, SP3, SP4 or SP5 polynucleotide sequences which may be used in accordance with the invention include deletions, insertions or substitutions of different nucleotide residues resulting in a polynucleotide that encodes the same or a functionally equivalent SP1, SP2, SP3, SP4 or SP5 homolog, respectively. As used herein a "deletion" is defined as a change in either nucleotide or amino acid sequence in which one or more nucleotides or amino acid residues, respectively, are absent.

As used herein an "insertion" or "addition" is that change in a nucleotide or amino acid sequence which has resulted in the addition of one or more nucleotides or amino acid residues, respectively, as compared to the naturally occurring SP1, SP2, SP3, SP4 or SP5.

As used herein "substitution" results from the replacement of one or more nucleotides or amino acids by different nucleotides or amino acids, respectively.

The encoded protein may also show deletions, insertions or substitutions of amino acid residues which produce a silent change and result in a functionally SP1, SP2, SP3, SP4 or SP5 variant. Deliberate amino acid substitutions may be made on the basis of similarity in polarity, charge, solubility, hydrophobicity, hydrophilicity, and/or the amphipathic nature of the residues as long as the variant retains the ability to modulate secretion. For example, negatively charged amino acids include aspartic acid and glutamic acid; positively charged amino acids include lysine and arginine; and amino acids with uncharged polar head groups having similar hydrophilicity values include leucine, isoleucine, valine; glycine, alanine; asparagine, glutamine; serine, threonine, phenylalanine, and tyrosine.

The SP1, SP2, SP3, SP4 or SP5 polynucleotides of the present invention may be engineered in order to modify the cloning, processing and/or expression of the gene product. For example, mutations may be introduced using techniques which are well known in the art, eg, site-directed mutagenesis to insert new restriction sites, to alter glycosylation patterns or to change codon preference, for example.

In one embodiment of the present invention, a gram-positive microorganism SP1, SP2, SP3, SP4 or SP5 polynucleotide may be ligated to a heterologous sequence to encode a fusion protein. A fusion protein may also be engineered to contain a cleavage site located between the serine protease nucleotide sequence and the heterologous protein sequence, so that the serine protease may be cleaved and purified away from the heterologous moiety.

IV. Vector Sequences

Expression vectors used in expressing the serine proteases of the present invention in gram-positive microorganisms comprise at least one promoter associated with a serine protease selected from the group consisting of SP1, SP2, SP3, SP4 and SP5, which promoter is functional in the host cell. In one embodiment of the present invention, the promoter is the wild-type promoter for the selected

serine protease and in another embodiment of the present invention, the promoter is heterologous to the serine protease, but still functional in the host cell. In one preferred embodiment of the present invention, nucleic acid encoding the serine protease is stably integrated into the microorganism genome.

In a preferred embodiment, the expression vector contains a multiple cloning site cassette which preferably comprises at least one restriction endonuclease site unique to the vector, to facilitate ease of nucleic acid manipulation. In a preferred embodiment, the vector also comprises one or more selectable markers. As used herein, the term selectable marker refers to a gene capable of expression in the gram-positive host which allows for ease of selection of those hosts containing the vector. Examples of such selectable markers include but are not limited to antibiotics, such as, erythromycin, actinomycin, chloramphenicol and tetracycline.

V. Transformation

A variety of host cells can be used for the production of SP1, SP2, SP3, SP4 or SP5 including bacterial, fungal, mammalian and insects cells. General transformation procedures are taught in Current Protocols In Molecular Biology (vol. 1, edited by Ausubel et al., John Wiley & Sons, Inc. 1987, Chapter 9) and include calcium phosphate methods, transformation using DEAE-Dextran and electroporation. Plant transformation methods are taught in Rodriquez (WO 95/14099, published 26 May 1995).

In a preferred embodiment, the host cell is a gram-positive microorganism and in another preferred embodiment, the host cell is *Bacillus*. In one embodiment of the present invention, nucleic acid encoding one or more serine protease(s) of the present invention is introduced into a host cell via an expression vector capable of replicating within the host cell. Suitable replicating plasmids for *Bacillus* are described in Molecular Biological Methods for *Bacillus*, Ed. Harwood and Cutting, John Wiley & Sons, 1990, hereby expressly incorporated by reference; see chapter 3 on plasmids. Suitable replicating plasmids for *B. subtilis* are listed on page 92.

In another embodiment, nucleic acid encoding a serine protease(s) of the present invention is stably integrated into the microorganism genome. Preferred host cells are gram-positive host cells. Another preferred host is *Bacillus*. Another preferred host is *Bacillus subtilis*. Several strategies have been described in the literature for the direct cloning of DNA in *Bacillus*. Plasmid marker rescue transformation involves the uptake of a donor plasmid by competent cells carrying a

partially homologous resident plasmid (Contente *et al.*, *Plasmid* 2:555-571 (1979); Haima *et al.*, *Mol. Gen. Genet.* 223:185-191 (1990); Weinrauch *et al.*, *J. Bacteriol. 154(3)*:1077-1087 (1983); and Weinrauch *et al.*, *J. Bacteriol. 169(3)*:1205-1211 (1987)). The incoming donor plasmid recombines with the homologous region of the resident "helper" plasmid in a process that mimics chromosomal transformation.

Transformation by protoplast transformation is described for *B. subtilis* in Chang and Cohen, (1979) Mol. Gen. Genet 168:111-115; for B.megaterium in Vorobjeva et al., (1980) FEMS Microbiol. Letters 7:261-263; for B. amyloliquefaciens in Smith et al., (1986) Appl. and Env. Microbiol. 51:634; for B.thuringiensis in Fisher et al., (1981) Arch. Microbiol. 139:213-217; for B.sphaericus in McDonald (1984) J. Gen. Microbiol. 130:203; and B.larvae in Bakhiet et al., (1985) 49:577. Mann et al., (1986, Current Microbiol. 13:131-135) report on transformation of *Bacillus* protoplasts and Holubova, (1985) Folia Microbiol. 30:97) disclose methods for introducing DNA into protoplasts using DNA containing liposomes.

VI. Identification of Transformants

Whether a host cell has been transformed with a mutated or a naturally occurring gene encoding a gram-positive SP1, SP2, SP3, SP4 or SP5, detection of the presence/absence of marker gene expression can suggests whether the gene of interest is present. However, its expression should be confirmed. For example, if the nucleic acid encoding a serine protease is inserted within a marker gene sequence, recombinant cells containing the insert can be identified by the absence of marker gene function. Alternatively, a marker gene can be placed in tandem with nucleic acid encoding the serine protease under the control of a single promoter. Expression of the marker gene in response to induction or selection usually indicates expression of the serine protease as well.

Alternatively, host cells which contain the coding sequence for a serine protease and express the protein may be identified by a variety of procedures known to those of skill in the art. These procedures include, but are not limited to, DNA-DNA or DNA-RNA hybridization and protein bioassay or immunoassay techniques which include membrane-based, solution-based, or chip-based technologies for the detection and/or quantification of the nucleic acid or protein.

The presence of the cysteine polynucleotide sequence can be detected by DNA-DNA or DNA-RNA hybridization or amplification using probes, portions or fragments of *B.subtilis* SP1, SP2, SP3, SP4 or SP5.

VII Assay of Protease Activity

There are various assays known to those of skill in the art for detecting and measuring protease activity. There are assays based upon the release of acid-soluble peptides from casein or hemoglobin measured as absorbance at 280 nm or colorimetrically using the Folin method (Bergmeyer, et al., 1984, Methods of Enzymatic Analysis vol. 5, Peptidases, Proteinases and their Inhibitors, Verlag Chemie, Weinheim). Other assays involve the solubilization of chromogenic substrates (Ward, 1983, Proteinases, in Microbial Enzymes and Biotechnology (W.M. Fogarty, ed.), Applied Science, London, pp. 251-317).

VIII. Secretion of Recombinant Proteins

Means for determining the levels of secretion of a heterologous or homologous protein in a gram-positive host cell and detecting secreted proteins include, using either polyclonal or monoclonal antibodies specific for the protein. Examples include enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA) and fluorescent activated cell sorting (FACS). These and other assays are described, among other places, in Hampton R et al (1990, Serological Methods, a Laboratory Manual, APS Press, St Paul MN) and Maddox DE et al (1983, J Exp Med 158:1211).

A wide variety of labels and conjugation techniques are known by those skilled in the art and can be used in various nucleic and amino acid assays. Means for producing labeled hybridization or PCR probes for detecting specific polynucleotide sequences include oligolabeling, nick translation, end-labeling or PCR amplification using a labeled nucleotide. Alternatively, the nucleotide sequence, or any portion of it, may be cloned into a vector for the production of an mRNA probe. Such vectors are known in the art, are commercially available, and may be used to synthesize RNA probes *in vitro* by addition of an appropriate RNA polymerase such as T7, T3 or SP6 and labeled nucleotides.

A number of companies such as Pharmacia Biotech (Piscataway NJ), Promega (Madison WI), and US Biochemical Corp (Cleveland OH) supply commercial kits and protocols for these procedures. Suitable reporter molecules or labels include those radionuclides, enzymes, fluorescent, chemiluminescent, or chromogenic agents as well as substrates, cofactors, inhibitors, magnetic particles and the like. Patents teaching the use of such labels include US Patents 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149 and 4,366,241. Also, recombinant immunoglobulins may be produced as shown in US Patent No. 4,816,567 and incorporated herein by reference.

IX Purification of Proteins

Gram positive host cells transformed with polynucleotide sequences encoding heterologous or homologous protein may be cultured under conditions suitable for the expression and recovery of the encoded protein from cell culture. The protein produced by a recombinant gram-positive host cell comprising a serine protease of the present invention will be secreted into the culture media. Other recombinant constructions may join the heterologous or homologous polynucleotide sequences to nucleotide sequence encoding a polypeptide domain which will facilitate purification of soluble proteins (Kroll DJ et al (1993) DNA Cell Biol 12:441-53).

Such purification facilitating domains include, but are not limited to, metal chelating peptides such as histidine-tryptophan modules that allow purification on immobilized metals (Porath J (1992) Protein Expr Purif 3:263-281), protein A domains that allow purification on immobilized immunoglobulin, and the domain utilized in the FLAGS extension/affinity purification system (Immunex Corp, Seattle WA). The inclusion of a cleavable linker sequence such as Factor XA or enterokinase (Invitrogen, San Diego CA) between the purification domain and the heterologous protein can be used to facilitate purification.

X. Uses of The Present Invention

Genetically Engineered Host Cells

The present invention provides genetically engineered host cells comprising preferably non-revertable mutations or deletions in the naturally occurring gene encoding one or more of SP1, SP2, SP3, SP4 or SP5 such that the proteolytic activity is diminished or deleted altogether. The host cell may contain additional protease deletions, such as deletions of the mature subtilisn protease and/or mature neutral protease disclosed in United States Patent No. 5,264,366.

In a preferred embodiment, the host cell is genetically engineered to produce a desired protein or polypeptide. In a preferred embodiment the host cell is a *Bacillus*. In another preferred embodiment, the host cell is a *Bacillus subtilis*.

In an alternative embodiment, a host cell is genetically engineered to produce a gram-positive SP1, SP2, SP3, SP4 or SP5. In a preferred embodiment, the host cell is grown under large scale fermentation conditions, the SP1, SP2, SP3, SP4 or

SP5 is isolated and/or purified and used in cleaning compositions such as detergents. WO 95/10615 discloses detergent formulation.

Polynucleotides

A B.subtlis SP1, SP2, SP3, SP4 or SP5 polynucleotide, or any part thereof, provides the basis for detecting the presence of gram-positive microorganism polynucleotide homologs through hybridization techniques and PCR technology.

Accordingly, one aspect of the present invention is to provide for nucleic acid hybridization and PCR probes which can be used to detect polynucleotide sequences, including genomic and cDNA sequences, encoding gram-positive SP1, SP2, SP3, SP4 or SP5 or portions thereof.

The manner and method of carrying out the present invention may be more fully understood by those of skill in the art by reference to the following examples, which examples are not intended in any manner to limit the scope of the present invention or of the claims directed thereto

Example I
Preparation of a Genomic library

The following example illustrates the preparation of a *Bacillus* genomic library.

Genomic DNA from *Bacillus* cells is prepared as taught in Current Protocols In Molecular Biology vol. 1, edited by Ausubel et al., John Wiley & Sons, Inc. 1987, chapter 2. 4.1. Generally, *Bacillus* cells from a saturated liquid culture are lysed and the proteins removed by digestion with proteinase K. Cell wall debris, polysaccharides, and remaining proteins are removed by selective precipitation with CTAB, and high molecular weight genomic DNA is recovered from the resulting supernatant by isopropanol precipitation. If exceptionally clean genomic DNA is desired, an additional step of purifying the *Bacillus* genomic DNA on a cesium chloride gradient is added.

After obtaining purified genomic DNA, the DNA is subjected to Sau3A digestion. Sau3A recognizes the 4 base pair site GATC and generates fragments compatible with several convenient phage lambda and cosmid vectors. The DNA is subjected to partial digestion to increase the chance of obtaining random fragments.

The partially digested *Bacillus* genomic DNA is subjected to size fractionation on a 1% agarose gel prior to cloning into a vector. Alternatively, size fractionation on a sucrose gradient can be used. The genomic DNA obtained from the size

fractionation step is purified away from the agarose and ligated into a cloning vector appropriate for use in a host cell and transformed into the host cell.

Example II

The following example describes the detection of gram-positive microorganism SP1. The same procedures can be used to detect SP2, SP3, SP4 or SP5.

DNA derived from a gram-positive microorganism is prepared according to the methods disclosed in Current Protocols in Molecular Biology, Chap. 2 or 3. The nucleic acid is subjected to hybridization and/or PCR amplification with a probe or primer derived from SP1. A preferred probe comprises the nucleic acid section encoding conserved amino acid residues.

The nucleic acid probe is labeled by combining 50 pmol of the nucleic acid and 250 mCi of [gamma ^{32}P] adenosine triphosphate (Amersham, Chicago IL) and T4 polynucleotide kinase (DuPont NEN®, Boston MA). The labeled probe is purified with Sephadex G-25 super fine resin column (Pharmacia). A portion containing 10^7 counts per minute of each is used in a typical membrane based hybridization analysis of nucleic acid sample of either genomic or cDNA origin.

The DNA sample which has been subjected to restriction endonuclease digestion is fractionated on a 0.7 percent agarose gel and transferred to nylon membranes (Nytran Plus, Schleicher & Schuell, Durham NH). Hybridization is carried out for 16 hours at 40 degrees C. To remove nonspecific signals, blots are sequentially washed at room temperature under increasingly stringent conditions up to 0.1 x saline sodium citrate and 0.5% sodium dodecyl sulfate. The blots are exposed to film for several hours, the film developed and hybridization patterns are compared visually to detect polynucleotide homologs of *B.subtilis* SP1. The homologs are subjected to confirmatory nucleic acid sequencing. Methods for nucleic acid sequencing are well known in the art. Conventional enzymatic methods employ DNA polymerase Klenow fragment, SEQUENASE® (US Biochemical Corp, Cleveland, OH) or Taq polymerase to extend DNA chains from an oligonucleotide primer annealed to the DNA template of interest.

Various other examples and modifications of the foregoing description and examples will be apparent to a person skilled in the art after reading the disclosure without departing from the spirit and scope of the invention, and it is intended that all such examples or modifications be included within the scope of the appended claims. All publications and patents referenced herein are hereby incorporated in their entirety.

4

CLAIM

We claim:

1. An isolated polynucleotide encoding SP2 from a gram positive microorganism.

2. The polynucleotide of Claim 1 wherein SP2 has the amino acid sequence shown in Figures 7A-7B.

3. An isolated SP2 encoding nucleic acid having the nucleic acid sequence as shown in Figures 7A-7B.

4. An isolated SP2 from a gram-positive microorganism.

5. The isolated SP2 of Claim 4 having the amino acid sequence as shown in Figure 7A-7B.

6. An isolated polynucleotide encoding SP3 from a gram positive microorganism.

7. The polynucleotide of Claim 6 wherein SP3 has the amino acid sequence shown in Figures 8A-8B.

8. The isolated SP3 encoding nucleic acid having the sequence as shown in Figure 8A-8B.

9. An isolated SP3 from a gram-positive microorganism.

10. The isolated SP3 of Claim 9 having the amino acid sequence as shown in Figures 8A-8B.

11. A gram-positive microorganism having a mutation or deletion of part or all of one or more of the genes encoding serine proteases selected from the group consisting of SP1, SP2, SP3, SP4 and SP5 said mutation or deletion resulting in the inactivation of the CP1 proteolytic activity.

12. The gram-positive microorganism according to Claims 11 that is a member of the family *Bacillus*.

13. The microorganism according to Claim 12 wherein the member is selected from the group consisting of B. licheniformis, B. lentus, B. brevis, B. stearothermophilus, B. alkalophilus, B. amyloliquefaciens, B. coagulans, B. circulans, B. lautus and *Bacillus* thuringiensis.

14. The microorganism of Claim 11 wherein said microorganism is capable of expressing a heterologous protein.

15. The microorganism of Claim 14 wherein said heterologous protein is selected from the group consisting of hormone, enzyme, growth factor and cytokine.

16. The host cell of Claim 15 wherein said heterologous protein is an enzyme.

17. The host cell of Claim 16 wherein said enzyme is selected from the group consisting of a proteases, carbohydrases, and lipases; isomerases such as racemases, epimerases, tautomerases, or mutases; transferases, kinases and phophatases.

18. A cleaning composition comprising a serine protease selected from the group consisting of SP1, SP2, SP3, SP4 and SP5.

19 An expression vector comprising nucleic acid encoding a serine protease selected from the group consisting of SP1, SP2, SP3, SP5 or SP5.

22. A host cell comprising an expression vector according to Claim 21.

ABSTRACT

The present invention relates to the identification of novel serine proteases in Gram-positive microorganisms. The present invention provides the nuclei acid and amino acid sequences for the *Bacillus subtilis* serine proteases SP1, SP2, SP3, SP4 and SP5. The present invention also provides host cells having a mutation or deletion of part or all of the gene encoding SP1, SP2, SP3, SP4 and SP5. The present invention also provides host cells further comprising nucleic acid encoding desired heterologous proteins such as enzymes. The present invention also provides a cleaning composition comprising a serine protease of the present invention.

```
         10                  30
atgaaaaagctgataaccgcagacgacatcacagcgattgtctctgtg
M  K  K  L  I  T  A  D  D  I  T  A  I  V  S  V

50                  70                  90
accgatcctcaatacgccccagacggtacccgtgccgcatatgtaaaa
T  D  P  Q  Y  A  P  D  G  T  R  A  A  Y  V  K

          110                 130
tcacaagtaaatcaagagaaagattcgtatacatcaaatatatggatc
S  Q  V  N  Q  E  K  D  S  Y  T  S  N  I  W  I

   150                 170                 190
tatgaaacgaaaacgggaggatctgttccttggacacatggagaaaag
Y  E  T  K  T  G  G  S  V  P  W  T  H  G  E  K

               210                 230
cgaagcaccgacccaagatggtctccggacgggcgcacgcttgccttt
R  S  T  D  P  R  W  S  P  D  G  R  T  L  A  F

       250                 270                 2
atttctgatcgagaaggcgatgcggcacagctttatatcatgagcact
I  S  D  R  E  G  D  A  A  Q  L  Y  I  M  S  T

90                  310                 330
gaaggcggagaagcaagaaaactgactgatatcccatatggcgtgtca
E  G  G  E  A  R  K  L  T  D  I  P  Y  G  V  S

          350                 370
aagccgctatggtccccggacggtgaatcgattctggtcactatcagt
K  P  L  W  S  P  D  G  E  S  I  L  V  T  I  S

   390                 410                 430
ttgggagagggggaaagcattgatgaccgagaaaaaacagagcaggac
L  G  E  G  E  S  I  D  D  R  E  K  T  E  Q  D

               450                 470
agctatgaacctgttgaagtgcaaggcctctcctacaaacgggacggc
S  Y  E  P  V  E  V  Q  G  L  S  Y  K  R  D  G

       490                 510                 5
aaagggctgacgagaggtgcgtatgcccagcttgtgcttgtcagcgta
K  G  L  T  R  G  A  Y  A  Q  L  V  L  V  S  V

30                  550                 570
aagtcgggtgagatgaaagagctgacaagtcacaaagctgatcatggt
K  S  G  E  M  K  E  L  T  S  H  K  A  D  H  G

          590                 610
gatcctgctttttctcctgacggcaaatggcttgttttctcagctaat
D  P  A  F  S  P  D  G  K  W  L  V  F  S  A  N

   630                 650                 670
ttaactgaaacagatgatgccagcaagccgcatgatgtttacataatg
L  T  E  T  D  D  A  S  K  P  H  D  V  Y  I  M

               690                 710
```

GC 382 Figure 1A

```
tcactggagtctggagatcttaagcaggttacacctcatcgcggctca
S  L  E  S  G  D  L  K  Q  V  T  P  H  R  G  S

       730                   750                   7
ttcggatcaagctcattttcaccagacggaaggtatcttgctttgctt
F  G  S  S  S  F  S  P  D  G  R  Y  L  A  L  L

70                   790                   810
ggaaatgaaaaggaatataagaatgctacgctctcaaaggcgtggctc
G  N  E  K  E  Y  K  N  A  T  L  S  K  A  W  L

          830                   850
tatgatatcgaacaaggccgcctcacatgtcttactgagatgctggac
Y  D  I  E  Q  G  R  L  T  C  L  T  E  M  L  D

   870                   890                   910
gttcatttagcggatgcgctgattggagattcattgatcggtggtgct
V  H  L  A  D  A  L  I  G  D  S  L  I  G  G  A

                930                   950
gaacagcgcccgatttggacaaaggacagccaagggttttatgtcatc
E  Q  R  P  I  W  T  K  D  S  Q  G  F  Y  V  I

       970                   990                  10
ggcacagatcaaggcagtacgggcatctattatatttcgattgaaggc
G  T  D  Q  G  S  T  G  I  Y  Y  I  S  I  E  G

10                1030                  1050
cttgtgtatccgattcgtctggaaaaagagtacatcaatagcttttct
L  V  Y  P  I  R  L  E  K  E  Y  I  N  S  F  S

          1070                  1090
ctttcacctgatgaacagcactttattgccagtgtgacaaagccggac
L  S  P  D  E  Q  H  F  I  A  S  V  T  K  P  D

  1110                  1130                  1150
agaccgagtgagctttacagtatcccgcttggacaggaagagaaacag
R  P  S  E  L  Y  S  I  P  L  G  Q  E  E  K  Q

               1170                  1190
ctgactggcgcgaatgacaagtttgtcagggagcatacgatatcaata
L  T  G  A  N  D  K  F  V  R  E  H  T  I  S  I

      1210                  1230                  12
cctgaagagattcaatatgctacagaagacggcgtgatggtgaacggc
P  E  E  I  Q  Y  A  T  E  D  G  V  M  V  N  G

50                1270                  1290
tggctgatgaggcctgcacaaatggaaggtgagacaacatatccactt
W  L  M  R  P  A  Q  M  E  G  E  T  T  Y  P  L

          1310                  1330
attcttaacatacacggcggtccgcatatgatgtacggacatacatat
I  L  N  I  H  G  G  P  H  M  M  Y  G  H  T  Y

  1350                  1370                  1390
tttcatgagtttcaggtgctggcggcgaaaggatacgcggtcgtttat
F  H  E  F  Q  V  L  A  A  K  G  Y  A  V  V  Y

               1410                  1430
```

6C382 Figure 1B

```
atcaatccgagaggaagccacggctacgggcaggaatttgtgaatgcg
I  N  P  R  G  S  H  G  Y  G  Q  E  F  V  N  A

      1450                 1470                 14
gtcagaggagattatgggggaaaggattatgacgatgtgatgcaggct
V  R  G  D  Y  G  G  K  D  Y  D  D  V  M  Q  A

90                1510                 1530
gtggatgaggctatcaaacgagatccgcatattgatcctaagcggctc
V  D  E  A  I  K  R  D  P  H  I  D  P  K  R  L

          1550                 1570
ggtgtcacgggcggaagctacggaggttttatgaccaactggatcgtc
G  V  T  G  G  S  Y  G  G  F  M  T  N  W  I  V

  1590                 1610                 1630
gggcagacgaaccgctttaaagctgccgttacccagcgctcgatatca
G  Q  T  N  R  F  K  A  A  V  T  Q  R  S  I  S

               1650                 1670
aattggatcagctttcacggcgtcagtgatatcggctatttctttaca
N  W  I  S  F  H  G  V  S  D  I  G  Y  F  F  T

      1690                 1710                 17
gactggcagcttgagcatgacatgtttgaggacacagaaaagctctgg
D  W  Q  L  E  H  D  M  F  E  D  T  E  K  L  W

30                1750                 1770
gaccggtctcctttaaaatacgcagcaaacgtggagacaccgcttttg
D  R  S  P  L  K  Y  A  A  N  V  E  T  P  L  L

          1790                 1810
atactgcatggcgagcgggatgaccgatgcccgatcgagcaggcggag
I  L  H  G  E  R  D  D  R  C  P  I  E  Q  A  E

  1830                 1850                 1870
cagctgtttatcgctctgaaaaaaatgggcaaggaaaccaagcttgtc
Q  L  F  I  A  L  K  K  M  G  K  E  T  K  L  V

               1890                 1910
cgttttccgaatgcatcgcacaatttatcacgcaccggacacccaaga
R  F  P  N  A  S  H  N  L  S  R  T  G  H  P  R

      1930                 1950                 19
cagcggatcaagcgcctgaattatatcagctcatggtttgatcaacat
Q  R  I  K  R  L  N  Y  I  S  S  W  F  D  Q  H

70
ctc
L
```

GC 382 Figure 1C

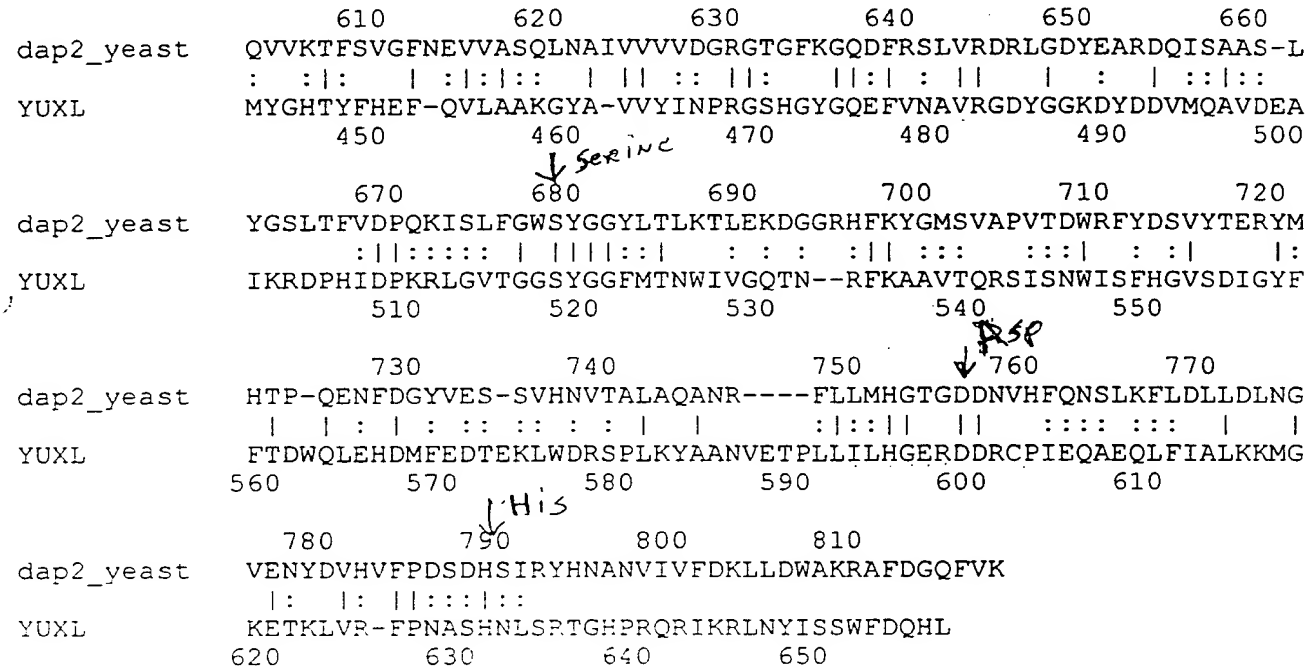
SCORES Init1: 58 Initn: 165 Opt: 204 z-score: 227.4 E(): 3e-07

Smith-Waterman score: 235; 20.3% identity in 646 aa overlap

```
                 170       180       190       200       210       220
dap2_yeast   WRHSTFGSYFVYDKSSSSFEEIGNEVALAIWSPNSNDIAYVQDN-NIYIYSAISKKTIRA
                                             ::|:::  |||::: |    |  |:  |
YUXL                        MKKLITADDITAIVSVTDPQYAPDGTRAAYVKSQVNQEKDSYTSNIWIYE
                                  10        20        30        40        50

                 230       240       250       260       270       280
dap2_yeast   VTNDGSSFLFNGKPDWVYEEEVFEDDKAAWWSPTGDYLAFLKIDESEVGEFIIPYYVQDE
             : : ||       | |:: |:   | :   ||| |  |||::  |:::::::   | :: |
YUXL         TKTGGSV------P-WTHGEKRSTDPR---WSPDGRTLAFISDREGDAAQL---YIMSTE
                           60        70            80        90

                 290           300       310       320           330
dap2_yeast   KDIYPEMRSIKYPKSG---TPNPHAELWVYSMKDGTSFHPRISGNKKDG--SLLITEVTW
             :: :| |  |    :|: :: | : |: :| |:  |   :::|:   : :  :::
YUXL         GGEARKLTDIPYGVSKPLWSPDGESILVTISLGEGESIDDR-EKTEQDSYEPVEVQGLSY
             100       110       120       130       140       150

             340       350       360       370           380       390
dap2_yeast   VGNGNVLVKTTDRSSDILTVFLIDTIAKTSNVVRNE----SSNGGWWEITHNTLFIPANE
               :|: |:: :  :  :::|   :    ||: : :     | :| |  :: |   :  ::
YUXL         KRDGKGLTRGAYAQLVLVSVKSGEMKELTSHKADHGDPAFSPDGKWLVFSAN---LTETD
              160       170       180       190       200           210

                 400       410           420           430       440
dap2_yeast   TFDRPHNGYVDILPIGGYN----HLAYFENSNSS---HYKTLTEGKWEVVNGPLA----F
               ::||: |:  |  |  :    | : | :|: |   :| :|  :: |  |: |:    :
YUXL         DASKPHDVYIMSLESGDLKQVTPHRGSFGSSSFSPDGRYLALLGNEKEYKNATLSKAWLY
                 220       230       240       250       260       270

                  450       460       470       480       490       499
dap2_yeast   DSMENRLYFISTRKSSTERHVYYID-LRSPNEIIEVTDTSEDGVYDVSFSSGRRFGL--L
             |  ::|| -::   :     :   | | :  |   :   : :| | :: ::|   |:  :
YUXL         DIEQGRLTCLTEMLDVHLADALIGDSLIGGAEQRPIWTKDSQGFYVIGTDQGST-GIYYI
                 280       290       300       310       320       330

           500       510       520       530       540       550
dap2_yeast   TYKGPKVPYQKIVDFHSRKAEKCDKGNVLGKSLYHLEKNEVLTKILEDYAVPR-KSFREL
             : :|   | :   :: :  : : |: : :::     : :|:       |::|  :  ::|
YUXL         SIEGLVYPIRLEKEYINSFSLSPDEQHFIASVTKPDRPSEL-------YSIPLGQEEKQL
                 340       350       360       370           380

            560                   570       580       590       600
dap2_yeast   NLGKDEFGKD----------------ILVNSYEILPNDFDETLSDHYPVFFFAYGGPNSQ
             : ::|:|  ::                ::|||:: : | :::   :  ||:::  :|||: :
YUXL         TGANDKFVREHTISIPEEIQYATEDGVMVNGWLMRPAQMEGETT--YPLILNIHGGPH-M
               390       400       410       420           430       440
```

GC382

Figure 2A



GC382

Figure 2B

SCORES Init1: 45 Initn: 45 Opt: 178 z-score: 198.3 E(): 1.2e-05
Smith-Waterman score: 215; 23.2% identity in 254 aa overlap

```
           380       390       400       410       420       430       439
yuxl.bsupep  QEEKQLTGANDKFVREHTISIPEEIQYATEDGVMVNGWLMRPAQMEGETTYPLILNIHGG
                                              :|: |:| | :||:  |:  |  :| ::||

YTMA                    MIVEKRRFPSPSQHVRLYTICYLSNGLRVKGLLAEPAE-PGQ--YDGFLYLRGG
                                10        20        30        40          50

           440       450       460       470       480       490
yuxl.bsupep  PHMMYGHTYFHEFQVLAAKGYAVVYINPRGSHG-YGQEFVNAVRGDYGGKDYDDVMQAVD
              : : | :   ::  :|:::|::|     ||::|  |:|       |::|:| :|:::|
YTMA         IKSV-GMVRPGRIIQFASQGFVVFAPFYRGNQGGEGNE-------DFAGEDREDAFSAF-
                       60        70        80                90        100

           500       510       520       530       540       550
yuxl.bsupep  EAIKRDPHIDPKRLGVTGGSYGGFMTNWIVGQTNRFKAAVTQRSISNWISFHGVSDIGYF
             : ::: |::   |: : | | ||:|     |: : :: :    |: :|    ||||:
YTMA         RLLQQHPNVKKDRIHIFGFSRGGIM-----GMLTAIEMGGQAASFVSW---GGVSDMILT
                  110       120 ↑SER      130       140           150

           560       570                   580       590       600
yuxl.bsupep  FTDWQLEHDMFEDT----------EKLWDRSPLKYAANVETPLLILHGERDDRCPIEQAE
             : : |  : |:: :           |  | |:|:  : ::::|:|::|||:|:   |:::
YTMA         YEERQDLRRMMKRVIGGTPKKVPEEYQW-RTPFDQVNKIQAPVLLIHGEKDQNVSIQHSY
                160       170       180       190       200  ↑  210
                                                             Asp

           610       620       630       640       650
yuxl.bsupep  QLFIALKKMGKETKLVRFPNASHNLSRTGHPRQRIKRLNYISSWFDQHL
              |   ||::_| ::   : : :| :     |::  : :  :::|:
YTMA         LLEEKLKQLHKPVETWYYSTFTHYFP----PKENRRIVRQLTQWMKNR
                220       230     ↑HIS 240       250
```

GC 382

Figure 3

```
SCORES      Initl:    58  Initn:    84  Opt:   153 z-score: 171.4 E(): 0.000
38
Smith-Waterman score: 153;    23.9% identity in 243 aa overlap

                    410       420       430       440       450       460
yuxl.bsupep  PEEIQYATEDGVMVNGWLMRPAQMEGETTYPLILNIHGGPHMMYGHTYFHEFQVLAAKGY
                                                ||:: |||       |: :|   :|| ||:
YITV                MIQIENQTVSGIPFLHIVKEENRHRAVPLVIFIHGFTSAKE-HN-LHIAYLLAEKGF
                           10        20        30        40          50

                    470       480       490       500               510
yuxl.bsupep  AVVYINPRGSHGYGQEFVNAVRGDYGGKDYDDVMQAVDEA------IKRDPHIDPKRLGV
              :|   |:: | :|::  : :  : :|: :| |:: ::|       ::::  ||  |:|:
YITV         RAVL--PEALH-HGERGEEMAVEELAGHFWDIVLNEIEEIGVLKNHFEKEGLIDGGRIGL
                  60        70        80        90       100       110

                 520       530       540       550       560       570
yuxl.bsupep  TGGSYGGFMTNWIVGQTNRFKAAVTQRSISNWISFHGVSDIGYFFTDWQLEHDMFED-TE
             :| |:||: |   :   : :||:|:  :  |:: :   ::| ::  :  :| |: |: ::
YITV         AGTSMGGITTLGALTAYDWIKAGVSLMGSPNYVELFQ-QQIDHI-QSQGIEIDVPEEKVQ
                ↑120       130       140       150       160       170
                SER
                 580              590       600       610       620
yuxl.bsupep  KLWDRSPLKYAANV-----ETPLLILHGERDDRCPIEQAEQLFIALKKMGKET----KLV
             :|  |  |:  :       : |||: || :|   |   ::::: ::|:  :|     :::
YITV         QLMKRLELRDLSLQPEKLQQRPLLFWHGAKDKVVPYAPTRKFYDTIKSHYSEQPERLQFI
                    180       190       200↑      210       220       230
                                            Asp

                 630       640       650
yuxl.bsupep  RFPNASHNLSRTGHPRQRIKRLNYISSWFDQHL
                ||:|::     ||  :  |: |  ||: :|
YITV         GDENADHKV-----PRAAV--LKTIE-WFETYL
                ↑His 240          250
```

Figure 4

GC 382

```
SCORES      Init1:     67  Initn:     67  Opt:   117 z-score: 131.5 E():  0.06
5
Smith-Waterman score: 117;     21.6% identity in 232 aa overlap

                390        400        410        420        430        440
yuxl.bsupep  TGANDKFVREHTISIPEEIQYATEDGVMVNGWLMRPAQMEGETTYPLILNIHGGP-HMMY

YQKD                                                  |: ::  ::  : |   |:  ||   :::
                IIKRETDNGHDVFESFEQMEKTAFVIPSAYGYDIKGYHVAPHDTPNTIIICHGVTMNVLN
                  40         50         60         70         80         90

                450        460        470        480        490        500
yuxl.bsupep  GHTYFHEFQVLAAKGYAVVYINPRGSHGYGQEFVNAVRGDYGGKDYDDVMQAVDEAIKRD
             :  |:| |  |    |: |:  : |  || :    ::    :||  : ||: ::|:   ::
YQKD         SLKYMHLFLDL---GWNVLIYDHR-RHGQS----GGKTTSYGFYEKDDLNKVVSLLKNKT
              100          110        120          130        140

                510        520        530          540        550       559
yuxl.bsupep  PHIDPKRLGVTGGSYGGFMTNWIVGQ-----TNRFKAAVTQRSISNWISFHGVSDIGYFF
             |     :|: | |:|:  :   :|      :: : |      ::: :::: :: | :
YQKD         NHRG--LIGIHGESMGAVTALLYAGAHCSDGADFYIADCPFACFDEQLAYRLRAE--YRL
            150          160 Serine 170        180        190        200

           560        570        580        590        600        610
yuxl.bsupep  TDWQLEH--DMFEDTE---KLWDRSPLKYAANVETPLLILHGERDDRCPIEQAEQLFIAL
              :| |    |:|   :   :  : |||    ::| |:|::|:: ||  |: ::|:|:
YQKD         PSWPLLPIADFFLKLRGGYRAREVSPLAVIDKIEKPVLFIHSKDDDYIPVSSTERLY--E
                210        220        230        240        250 Asp  260

                620        630        640        650
yuxl.bsupep  KKMGKETKLVRFPNASHNLSRTGHPRQRIKRLNYISSWFDQHL
             || | ::  :   |: | :| |
                                 His
YQKD         KKRGPKALYIA-ENGEHAMSYTKNRHTYRKTVQEFLDNMNDSTE
              270        280        290        300
```

GC 382

Figure 5

```
SCORES      Init1:    66  Initn:    90  Opt:   114 z-score: 128.0 E():     0.
1
Smith-Waterman score: 152;    25.1% identity in 303 aa overlap

                   330       340       350       360       370       379
yuxl.bsupep  GTDQGSTGIYYISIEGLVYPIRLEKEYINSFSLSPDE-QHFIASVTKPDRPSELYSIPLG
                                          |:|  |: | :    | |   ||:::: |
CAH                                      MQLFDLPLDQLQTYKPEKTAPKDFSEFWKLSLE
                                                10        20        30

           380       390       400       410       420       430
yuxl.bsupep  QEEKQLTGANDKFVREHTISIP-EEIQYATEDGVMVNGWLMRPAQMEGETTYPLILNIHG
             :  |  :  : : |   : ::   :: | :  :: ::||   | : ||   :| |:: ||
CAH          ELAKVQAEPDLQPVDYPADGVKVYRLTYKSFGNARITGWYAVPDK-EGP--HPAIVKYHG
                  40        50        60        70        80          90

           440       450       460                   470       480
yuxl.bsupep  GPHMMYGHTYFHEFQVLAAKGYAV-------------VYINPRGSHGYGQEFVNAVRGD-
               : |:  :||:   | :|||:              : |:|:| |: |    : :  |
CAH          YNASYDGE--IHEMVNWALHGYATFGMLVRGQQSSEDTSISPHG-HALGWMTKGILDKDT
                    100       110       120       130       140

                 490       500       510       520       530       540
yuxl.bsupep  --YGGKDYDDVMQAVDEAIKRDPHIDPKRLGVTGGSYGGFMTNWIVGQTNRFKAAVTQRS
               | |  | |:::|: |:|:   ::|  |:|||||| || :|   :: ::  ||||::
CAH          YYYRGV-YLDAVRAL-EVISSFDEVDETRIGVTGGSQGGGLTIAAAALSDIPKAAVADYP
             150       160       170       180       190       200
                                                  ↑SER

                   550               560       570       580       590
yuxl.bsupep  -ISNWISFHGVS------DIGYFFTDWQLEHDMFEDTEKLWDRSPLKYAANVETPLLILH
              :||:    |:      :|: ||      :   :  : |   : :: |  |::|:|:
CAH          YLSNFERAIDVALEQPYLEINSFFRRNGSPETEVQAMKTLSYFDIMNLADRVKVPVLMSI
               210       220       230       240       250       260

               600       610       620       630       640       650
yuxl.bsupep  GERDDRCPIEQAEQLFIALKKM--GKETKLVRFPNASHNLSRTGHPRQRIKRLNYISSWF
             |  |   |      :| | :::   || |: |:
CAH          GLIDKVTP---PSTVFAAYNHLETKKELKVYRYFGHEYIPAFQTEKLAFFKQHLKG
               270           280       290       300       310
                                                  ↑HIS
               ↑ASP
```

Figure 6

GC 382

```
        10                  30
ttgattgtagagaaaagaagatttccgtcgccaagccagcatgtgcgt
L  I  V  E  K  R  R  F  P  S  P  S  Q  H  V  R

50                  70                  90
ttgtatacgatctgctatctgtcaaatggattacgggttaaggggctt
L  Y  T  I  C  Y  L  S  N  G  L  R  V  K  G  L

           110                 130
ctggctgagccggcggaaccgggacaatatgacggatttttatatttg
L  A  E  P  A  E  P  G  Q  Y  D  G  F  L  Y  L

   150                 170                 190
cgcggcgggattaaaagcgtgggcatggttcggccgggccggattatc
R  G  G  I  K  S  V  G  M  V  R  P  G  R  I  I

                210                 230
cagtttgcatcccaagggtttgtggtgtttgctccttttacagaggc
Q  F  A  S  Q  G  F  V  V  F  A  P  F  Y  R  G

        250                 270                 2
aatcaaggaggagaaggcaatgaggattttgccggagaagacagggag
N  Q  G  G  E  G  N  E  D  F  A  G  E  D  R  E

90                  310                 330
gatgcattttctgcttttcgcctgcttcagcagcacccaaatgtcaag
D  A  F  S  A  F  R  L  L  Q  Q  H  P  N  V  K

           350                 370
aaggatagaatccatatcttcggtttttcccgcggcggaattatggga
K  D  R  I  H  I  F  G  F  S  R  G  G  I  M  G

   390                 410                 430
atgctcactgcgatcgaaatgggcgggcaggcagcttcatttgtttcc
M  L  T  A  I  E  M  G  G  Q  A  A  S  F  V  S

                450                 470
tggggaggcgtcagtgatatgattcttacatacgaggagcggcaggat
W  G  G  V  S  D  M  I  L  T  Y  E  E  R  Q  D

        490                 510                 5
ttgcggcgaatgatgaaaagagtcatcggcggaacaccgaaaaaggtg
L  R  R  M  M  K  R  V  I  G  G  T  P  K  K  V

30                  550                 570
cctgaggaatatcaatggaggacaccgtttgaccaagtaaacaaaatt
P  E  E  Y  Q  W  R  T  P  F  D  Q  V  N  K  I

           590                 610
caggctcccgtgctgttaatccatggagaaaaagaccaaaatgtttcg
Q  A  P  V  L  L  I  H  G  E  K  D  Q  N  V  S

   630                 650                 670
attcagcattcctatttattagaagagaagctaaaacaactgcataag
I  Q  H  S  Y  L  L  E  E  K  L  K  Q  L  H  K

        690                 710
```

GC 382 Figure 7A

```
ccggtggaaacatggtactacagtacattcacacattatttcccgcca
P  V  E  T  W  Y  Y  S  T  F  T  H  Y  F  P  P

       730                   750                  7
aaagaaaaccggcgtatcgtgcggcagctcacacaatggatgaaaaac
K  E  N  R  R  I  V  R  Q  L  T  Q  W  M  K  N

70
cgc
R
```

GC 382 Figure 7B

```
        10                  30
gtgatacaaattgagaatcaaaccgtttccggtattccgtttttacat
V I Q I E N Q T V S G I P F L H

50                  70                  90
attgtaaaggaagagaacaggcaccgcgctgttcctctcgtgatcttt
I V K E E N R H R A V P L V I F

          110                 130
atacatggttttacaagcgcgaaggaacacaaccttcatattgcttat
I H G F T S A K E H N L H I A Y

   150                 170                 190
ctgcttgcggagaagggttttagagccgttctgccggaggctttgcac
L L A E K G F R A V L P E A L H

               210                 230
catggggaacggggagaagaaatggctgttgaagagctggcggggcat
H G E R G E E M A V E E L A G H

       250                 270                 2
ttttgggatatcgtcctcaacgagattgaagagatcggcgtacttaaa
F W D I V L N E I E E I G V L K

90              310                 330
aaccattttgaaaaagagggcctgatagacggcggccgcatcggtctc
N H F E K E G L I D G G R I G L

          350                 370
gcaggcacgtcaatgggcggcatcacaacgcttggcgctttgactgca
A G T S M G G I T T L G A L T A

   390                 410                 430
tatgattggataaaagccggcgtcagcctgatgggaagcccgaattac
Y D W I K A G V S L M G S P N Y

               450                 470
gtggagctgtttcagcagcagattgaccatattcaatctcagggcatt
V E L F Q Q Q I D H I Q S Q G I

       490                 510                 5
gaaatcgatgtgccggaagagaaggtacagcagctgatgaaacgtctc
E I D V P E E K V Q Q L M K R L

30                  550                 570
gagttgcgggatctcagccttcagccggagaaactgcaacagcgcccg
E L R D L S L Q P E K L Q Q R P

          590                 610
cttttattttggcacggcgcaaaagataaagttgtgccttacgcgccg
L L F W H G A K D K V V P Y A P

   630                 650                 670
acccggaaattttatgacacgattaaatcccattacagcgagcagccg
T R K F Y D T I K S H Y S E Q P

               690                 710
```

GC 382 Figure 8A

gaacgcctgcaatttatcggagatgaaaacgctgaccataaagtcccg
E R L Q F I G D E N A D H K V P

730 750
cgggcagctgtgttaaaaacgattgaatggtttgaaacgtactta
R A A V L K T I E W F E T Y L

GC 382 Figure 8B

```
        10                  30
ttgaagaaaatccttttggccattggcgcgctcgtaacagctgtcatc
L  K  K  I  L  L  A  I  G  A  L  V  T  A  V  I

50                  70                  90
gcaatcggaattgtttttttcacatatgattctattcatcaagaaaaaa
A  I  G  I  V  F  S  H  M  I  L  F  I  K  K  K

           110                 130
acggatgaagacattatcaaaagagagacagacaacggacatgatgtg
T  D  E  D  I  I  K  R  E  T  D  N  G  H  D  V

   150                 170                 190
tttgaatcatttgaacaaatggagaaaaccgcttttgtgataccctcc
F  E  S  F  E  Q  M  E  K  T  A  F  V  I  P  S

                210                 230
gcttacgggtacgacataaaaggataccatgtcgcaccgcatgacaca
A  Y  G  Y  D  I  K  G  Y  H  V  A  P  H  D  T

       250                 270                  2
ccaaataccatcatcatctgccacggggtgacgatgaatgtactgaat
P  N  T  I  I  I  C  H  G  V  T  M  N  V  L  N

90                  310                 330
tctcttaagtatatgcatttatttctagatctcggctggaatgtgctc
S  L  K  Y  M  H  L  F  L  D  L  G  W  N  V  L

           350                 370
atttatgaccatcgccggcatggccaaagcggcggaaagacgaccagc
I  Y  D  H  R  R  H  G  Q  S  G  G  K  T  T  S

   390                 410                 430
tacgggttttacgaaaaggatgatctcaataaggttgtcagcttgctc
Y  G  F  Y  E  K  D  D  L  N  K  V  V  S  L  L

                450                 470
aaaaacaaaacaaatcatcgcggattgatcggaattcatggtgagtcg
K  N  K  T  N  H  R  G  L  I  G  I  H  G  E  S

       490                 510                  5
atgggggccgtgaccgccctgctttatgctggtgcacactgcagcgat
M  G  A  V  T  A  L  L  Y  A  G  A  H  C  S  D

30                  550                 570
ggcgctgatttttatattgccgattgtccgttcgcatgttttgatgaa
G  A  D  F  Y  I  A  D  C  P  F  A  C  F  D  E

           590                 610
cagcttgcctatcggctgagagcggaatacaggctcccgtcttggccc
Q  L  A  Y  R  L  R  A  E  Y  R  L  P  S  W  P

   630                 650                 670
ctgcttcctatcgccgacttctttttgaagctgaggggaggctatcgc
L  L  P  I  A  D  F  F  L  K  L  R  G  G  Y  R

                690                 710
```

GC 382 Figure 9A

```
gcacgtgaagtatctccgcttgctgtcattgataaaattgaaaagccg
A  R  E  V  S  P  L  A  V  I  D  K  I  E  K  P

       730                   750                  7
gtcctctttattcacagtaaggatgatgactacattcctgtttcttca
V  L  F  I  H  S  K  D  D  D  Y  I  P  V  S  S

70                   790                   810
accgagcggctttatgaaaagaaacgcggtccgaaagcgctgtacatt
T  E  R  L  Y  E  K  K  R  G  P  K  A  L  Y  I

            830                   850
gccgagaacggtgaacacgccatgtcatataccaaaaatcggcatacg
A  E  N  G  E  H  A  M  S  Y  T  K  N  R  H  T

   870                   890                   910
taccgaaaaacagtgcaggagtttttagacaacatgaatgattcaaca
Y  R  K  T  V  Q  E  F  L  D  N  M  N  D  S  T


gaa
E
```

GC 382 Figure 9B